Interpreting IV
What Economic Questions Can LATE Answer?

James J. Heckman
University of Chicago
Extract from: Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy (JEL 2010)

Econ 312, Spring 2023
This draft, March 30, 2023 12:38 Noon

THE UNIVERSITY OF CHICAGO

Making Explicit the Implicit Economics of LATE

- Vytlacil (2002, *Econometrica*): LATE is **equivalent** to a nonparametric version of the Generalized Roy model.
- The Imbens-Angrist conditions imply the generalized Roy model, and the generalized Roy model implies the LATE model.

- Vytlacil's analysis clarifies the implicit economic assumptions of LATE, what features of the generalized Roy model LATE estimates, and what policy questions LATE addresses.
- It also extends the range of policy questions that LATE can answer.

- By Vytlacil's theorem, the Imbens-Angrist conditions imply (and are implied by) a continuous latent variable discrete choice model, which represents the individual's decision to enroll in the program being studied.

- Treatment choice equation underlying LATE can be expressed in terms of observed (Z) and unobserved (V) variables that can be represented by: $I_D = \mu_D(Z) - V$ and $D = 1$ if $I_D > 0$; $D = 0$ otherwise, where V is a continuous random variable with distribution function F_V.
- V may depend on U_0 and U_1 in a general way.
- The counterfactual choice indicator is generated by: $D(z) = \mathbf{1}(\mu_D(z) > V)$.
- Additive separability between $\mu_D(Z)$ and V plays an essential role in LATE.
- $\underbrace{\mu_D(Z)}_{I_D} > V$ is an index.

- $P(z) \equiv \Pr(D = 1|Z = z)$.
- $P(z) = \Pr(\mu_D(z) > V) = F_V(\mu_D(z))$: scale transform of utility as a function of observed Z.

- Define: $U_D = F_V(V)$
- Uniformly distributed over the interval $[0, 1]$
- p^{th} quantile of U_D is p, i.e., the proportion of U_D that is p or lower.

$$D = \mathbf{1}(P(Z) > U_D). \tag{1}$$

- From $P(z)$, one can identify the *ex-ante* net benefit I_D up to scale and determine for each value of $Z = z$, what proportion of people perceive that they will benefit from the program and the intensity of their benefit.
- From agent choices, one can supplement the information in LATE and ascertain *ex-ante* subjective evaluations.

- The LATE assumptions imply the selection model representation (i.e., the generalized Roy model) and using the selection model representation, one can establish that $E(Y \mid Z = z) = E(Y \mid P(Z) = P(z))$: **index sufficiency**.

- As a consequence of Vytlacil's theorem, one can define $\text{LATE}(z^2, z^1)$ using the latent variable U_D and the values taken by $P(Z)$ when $Z = z^1$ and $Z = z^2$.
- Use the property that the Z enter the model only through $P(Z)$.

$$\text{LATE}(z^2, z^1) = E(Y_1 - Y_0 \mid P(z^1) \leq U_D \leq P(z^2)). \qquad (2)$$

- The mean gross return to persons whose $U_D \in [P(z^1), P(z^2)]$.

- LATE can be defined within the generalized Roy model, without reference to an instrument.
- LATE produced by economic theory can be expressed as

$$\text{LATE}(\bar{u}_D, \underline{u}_D) = E(Y_1 - Y_0 \mid \underline{u}_D \leq U_D \leq \bar{u}_D), \tag{3}$$

  the mean gross return to persons whose $U_D \in [\underline{u}_D, \bar{u}_D]$.
- A choice of two values of Z (z^1 and z^2) picks specific values of $[\underline{u}_D, \bar{u}_D]$ that identify the model-generated LATE from data (say $\Pr(D = 1 \mid Z = z^1) = p_1 = \underline{u}_D$ and $\Pr(D = 1 \mid Z = z^2) = p_2 = \bar{u}_D$).

The Surplus From Treatment and the Marginal Treatment Effect

- Using Vytlacil's theorem, it is possible to understand more deeply what economic questions LATE answers.
- For $P(Z) = p$, the mean gross gain of moving from "0" to "1" for people with U_D less than or equal to p is

$$\begin{aligned} E(Y_1 - Y_0 \mid P(Z) \geq U_D, P(Z) = p) \\ = E(Y_1 - Y_0 \mid p \geq U_D) \\ = E(Y_1 - Y_0 \mid \mu_D(z) \geq V). \end{aligned} \tag{4}$$

- The mean gross gain in the population (or gross surplus $S(p)$) that arises from participation in the program for people whose U_D is at or below p and the proportion of people whose U_D is at or below p: $E(Y_1 - Y_0 \mid p \geq U_D)p = S(p)$.

$$
\begin{aligned}
E(Y \mid P(Z) = p) &= E(Y_0 + \mathbf{1}(p \geq U_D)(Y_1 - Y_0)) \qquad (5) \\
&= E(Y_0) + \underbrace{E(Y_1 - Y_0 \mid p \geq U_D)p}_{S(p)}.
\end{aligned}
$$

- Can identify the left-hand side of (5) for all values of p in the support of $P(Z)$.

- It is not necessary to impose functional forms to obtain this expression, and one can avoid one of the common criticisms directed against structural econometrics.
- The surplus can be **defined** for all values of $p \in [0, 1]$ whether or not the model is identified.

- Marginal increment in outcomes is

$$\frac{\partial E(Y \mid P(Z)=p)}{\partial p} = \underbrace{E(Y_1 - Y_0 \mid U_D = p)}_{\text{MTE}} = \frac{\partial S(p)}{\partial p}. \quad (6)$$

- The sample analogue of (6) is the local instrumental variable (LIV) estimator of Heckman and Vytlacil (1999, 2005).
- Adopting a nonparametric approach to estimating $E(Y \mid P(Z)=p)$ avoids extrapolation outside of the sample support of $P(Z)$ and produces a data sensitive structural analysis.

- A generalization of this parameter defined for other points of evaluation of u_D is the Marginal Treatment Effect (MTE) is

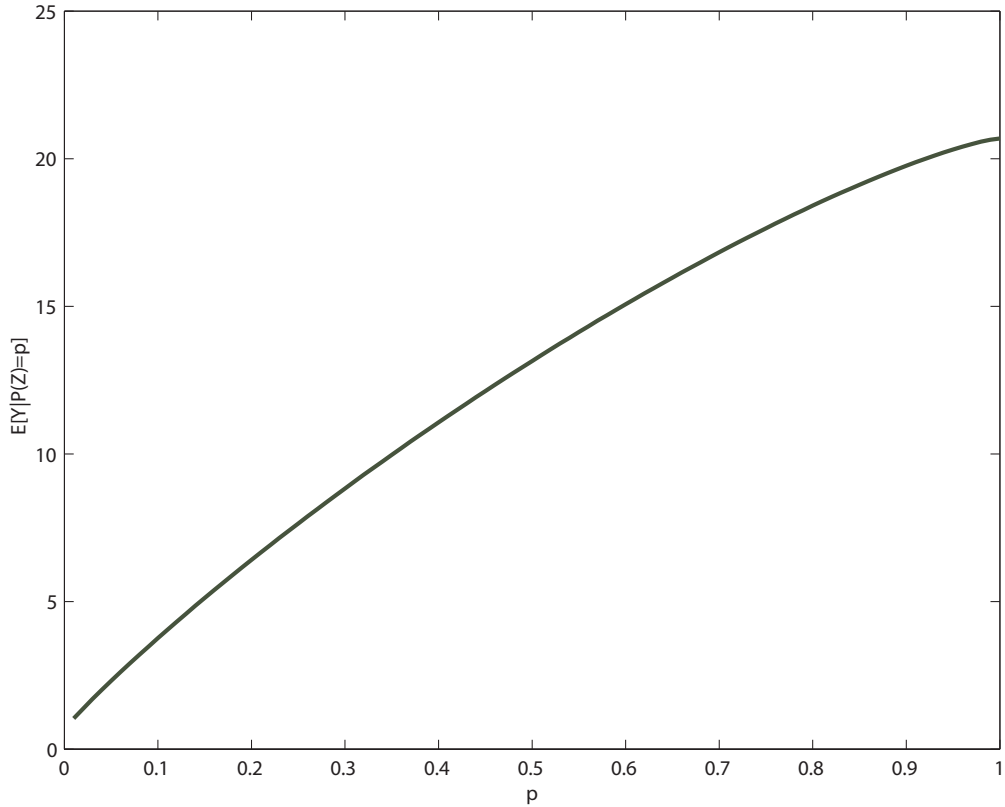
$$\text{MTE}(u_D) \equiv E(Y_1 - Y_0 \mid U_D = u_D).$$

- Expression (5) can be simplified to

$$E\left(Y \mid P(Z) = p\right) = E(Y_0) + \underbrace{\int_0^p \text{MTE}(u_D)\, du_D}_{S(p)}, \tag{7}$$

$$\frac{\partial S(p)}{\partial p} = \text{MTE}(p).$$

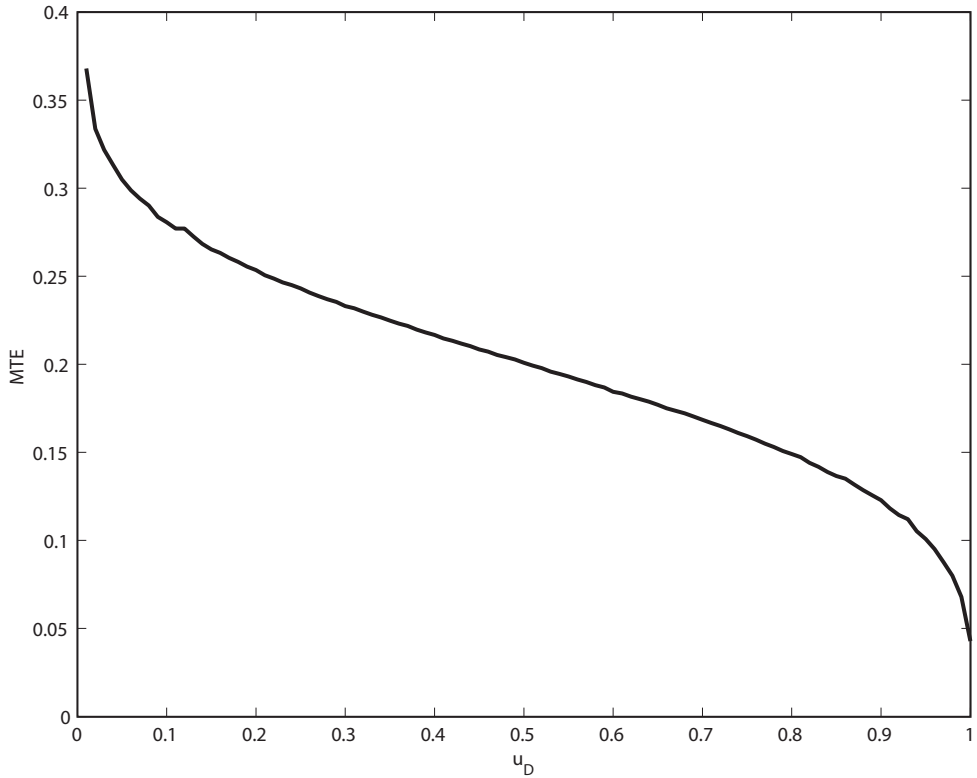
Figure 1: Plots of $E(Y|P(Z) = p)$ and the MTE derived from $E(Y|P(Z) = p)$



Plot of the $E(Y|P(Z) = p$
Source: Heckman and Vytlacil (2005).

THE UNIVERSITY OF CHICAGO

Figure 1: Plots of $E(Y|P(Z) = p)$ and the MTE derived from $E(Y|P(Z) = p)$



Plot of MTE(u_D): The derivative of $E(Y \mid P(Z) = p)$ evaluated at points $p = u_D$
Source: Heckman and Vytlacil (2005).

- From LIV, it is possible to identify returns at all quantiles of U_D within the support of the distribution of $P(Z)$ to determine which persons (identified by the quantile of the unobserved component of the desire to go to college, U_D) are induced to go into college ($D = 1$) by a marginal change in $P(z)$, i.e., analysts can define the margins of choice traced out by variations in different instruments as they shift $P(z)$.
- This clarifies what empirical versions of LATE identify by showing that all instruments operate through $P(Z)$, and variations around different levels of $P(Z)$ identify different stretches of the MTE.

The Fundamental Role of the Choice Probability in Understanding What Instrumental Variables Estimate When β Sorted by D

- For $p_2 > p_1$,

$$\begin{aligned} S(p_2) - S(p_1) &= E(Y_1 - Y_0 \mid p_1 \leq U_D \leq p_2) \Pr(p_1 \leq U_D \leq p_2) \\ &= E(Y_1 - Y_0 \mid p_1 \leq U_D \leq p_2)(p_2 - p_1), \end{aligned}$$

  note that $\Pr(p_1 \leq U_D \leq p_2) = p_2 - p_1$.

- Thus,

$$S(p_2) - S(p_1) = \int_{p_1}^{p_2} \text{MTE}(u_D)\, du_D.$$

$$\text{LATE}(p_2, p_1) = \frac{\int_{p_1}^{p_2} \text{MTE}(u_D)\, du_D}{p_2 - p_1} = \frac{S(p_2) - S(p_1)}{p_2 - p_1}. \tag{8}$$

- By the mean value theorem, $\text{LATE}(p_2, p_1) = \text{MTE}(u_D(p_2, p_1))$ where $u_D(p_2, p_1)$ is a point of evaluation and $u_D(p_2, p_1) \in [p_1, p_2]$.

- The model-generated LATE can be identified if there are values of Z , say $\tilde{z}$ and $\tilde{\tilde{z}}$, such that $\Pr(D = 1 \mid Z = \tilde{z}) = p_1$ and $\Pr(D = 1 \mid Z = \tilde{\tilde{z}}) = p_2$.
- Under standard regularity conditions

$$\lim_{p_2 \to p_1} \text{LATE}(p_2, p_1) = \text{MTE}(p_1).$$

- Partition the support of u_D into M discrete and exhaustive intervals

$$[u_{D,0}, u_{D,1}), [u_{D,1}, u_{D,2}), \ldots, [u_{D,M-1}, u_{D,M}],$$

where $u_{D,0} = 0$ and $u_{D,M} = 1$,

$$E(Y \mid U_D \leq u_{D,k}) = E(Y_0) + \sum_{j=1}^{k} \text{LATE}(u_{D,j}, u_{D,j-1})\eta_j,$$

where $\eta_j = u_{D,j} - u_{D,j-1}$.

- Thus

$$E(Y) = E(Y_0) + \sum_{j=1}^{M} \text{LATE}(u_{D,j}, u_{D,j-1})\eta_j. \tag{9}$$

- Counterpart to expression (7).
- It shows how mean income can be represented as a sum of incremental gross surpluses above $E(Y_0)$.

- If $\Pr(D = 1 \mid Z = z)$ assumes values at only a discrete set of support points, say $p_1 < p_2 < \cdots < p_L$, we can only identify LATE in intervals with boundaries defined by
  $u_{D,\ell} = p_\ell, \quad \ell = 1, \ldots, L.$

- MTE(u_D) and the model-generated LATE (2) are structural parameters in the sense that changes in Z (conditional on X) do not affect MTE(u_D) or theoretical LATE.
- They are invariant with respect to all policy changes that operate through Z.
- Conditional on X, one can transport MTE and the derived theoretical LATEs across different policy environments and different data sets.
- These policy invariant parameters implement Marschak's Maxim since they are defined for combinations of the parameters of the generalized Roy model.

- This deeper understanding of LATE facilitates its use in answering out of sample policy questions **P2** and **P3** for policies that operate through changing Z.
- Thus if one computes a LATE for any two pairs of values $Z = z^1$, and $Z = z^2$, with associated probabilities $\Pr(D = 1 \mid Z = z^1) = P(z^1) = p_1$ and $\Pr(D = 1 \mid Z = z^2) = P(z^2) = p_2$, one can use it to evaluate any other pair of policies $\tilde{z}$ and $\tilde{\tilde{z}}$ such that

$$\Pr(D = 1 \mid Z = z^1) = \Pr(D = 1 \mid Z = \tilde{z}) = p_1$$

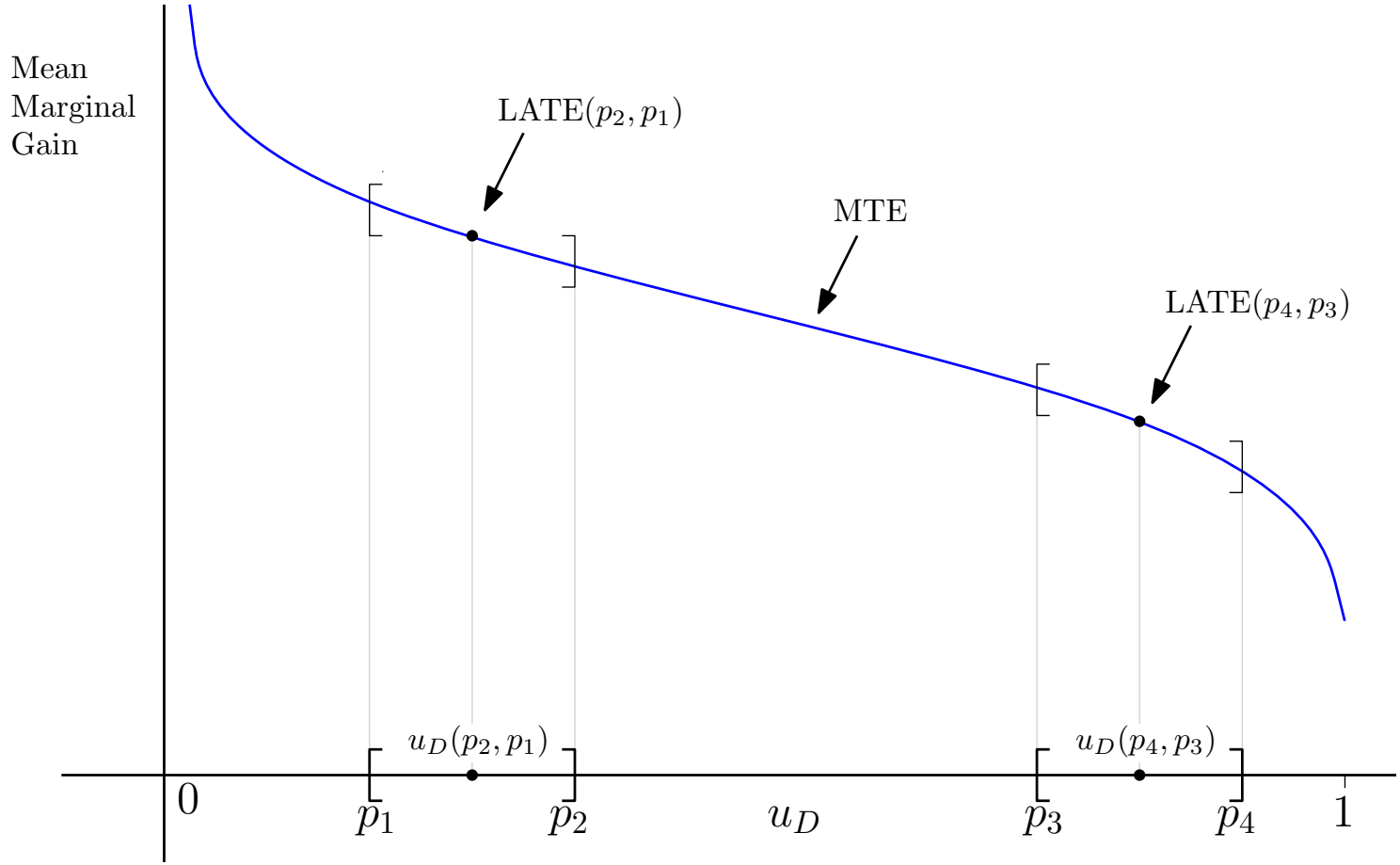
and

$$\Pr(D = 1 \mid Z = z^2) = \Pr(D = 1 \mid Z = \tilde{\tilde{z}}) = p_2.$$

- Thus, one can use an empirical LATE determined for one set of instrument configurations to identify outcomes for other sets of instrument configurations that produce the same p_1 and p_2, i.e., we can compare any policy described by $\tilde{z} \in \{z \mid P(z) = p_1\}$ with any policy $\tilde{\tilde{z}} \in \{z \mid P(z) = p_2\}$ and not just the policies associated with z^1 and z^2 that identify the sample LATE.
- This is a useful result and enables analysts to solve policy evaluation question **P3** to evaluate new policies never previously implemented if they can be cast in terms of variations in $P(Z)$ over the empirical support on Z.

- Variation in different components of Z produce variation in $P(Z)$.
- Analysts can aggregate the variation in different components of Z into the induced variation in $P(Z)$ to trace out MTE(u_D) over more of the support of u_D than would be possible using variation in any particular component of Z.
- The structural approach enables analysts to determine what stretches of the MTE different instruments identify and to determine the margin of U_D identified by the variation in an instrument.

Figure 2: MTE as a function of u_D: What sections of the MTE different values of the instruments and different instruments approximate.



- Instruments associated with higher values of $P(Z)$, $[p_3, p_4]$, identify the LATE in a different stretch of the MTE associated with higher values of u_D.
- Continuous instruments identify entire stretches of the MTE while discrete instruments define the MTE at discrete points of the support (i.e., the LATE associated with the interval defined by the values assumed by $P(Z)$).

- If the MTE does not depend on u_D, $E(Y \mid P(Z) = p) = E(Y_0) + (\mu_1 - \mu_0)p$, and all instruments identify the same parameter: $\bar{\beta} = \mu_1 - \mu_0$.
- In this case, MTE is a flat line parallel to the u_D axis.

- A test of whether MTE(u_D) depends on u_D, or a test of nonlinearity of $E(Y \mid P(Z) = p)$ in p, is a test of the whether different instruments estimate the same parameter.
- The LATE model and its extensions overturn the logic of the Durbin (1954)–Wu (1973)–Hausman (1978) test for overidentification.
- Variability among the estimates from IV estimators based on different instruments may have nothing to do with the validity of any particular instrument, but may just depend on what stretch of the MTE they approximate.